Math 108, Fall 2003

Quantitative Methods for Public Policy

Daniel Kaplan, Macalester College

Notes on Detection

Problems of **detection** arise in many settings: medical screening and diagnosis, corporate or individual cheating, identification of faults in systems (e.g., design flaws leading to rollover of 15-passenger vans, airport security screening), and others.

A more general term than detection is **classification**. A classification system is composed of several parts:

An Object e.g., a person in a medical diagnosis problem, or the design of a van, or an airport screening procedure, etc.

A Set of States e.g., cancerous or not, faulty design or not, etc. These states are the *true* condition of the object.

A Classification Test which takes information from the object and gives as a result a label: which one of the states the object is in according to the test. This result may be different from the true state.

Some examples of classification tests: an HIV test gives a positive or negative result to indicate whether a person is infected with HIV; a study of accident records indicates whether a vehicle has a faulty design with respect to rollovers.

A detection problem is a classification problem with only two states. These states are often designated simply positive or negative, yes or no, or true or false. In these notes we will use + or −. To distinguish between the true state and the result of the classification test, we will use P and N as the symbols for the result of the test.

Detection involves an inference from the result of the classification test to the true state of the object. When the classification test is perfect — that is, it always gives the right result — then the inference is trivial. But even conventional tests in widespread use may be far from perfect; care is then needed in thinking about the results of a test and what they say about reality.

It is often the case that a perfect or near perfect test is available but an imperfect test is used for reasons of economy or practicability. For example, in HIV testing, the expensive Western Blot technique gives a definitive test result, but the inexpensive ELISA method is used for screening. In contrast, consider malignant breast cancer. Self-examination gives a simple, practical test, the relatively inexpensive mammography procedure gives a somewhat more comprehensive test, and biopsy gives a more definitive test, but there is no absolutely definitive test. (Even cause of death is not definitive, since people with lethal breast cancer may die from other causes.)

In a **performance demonstration** of a detection test, a common procedure is to gather information about a set of objects. This set of objects is called the **demonstration population**. For each object in the demonstration population the true state and the test result are collected and tabulated. Each object has either a + or − true state and a P or N test result. There are, therefore, four possibilities altogether for each object: $+P$ or $+N$ or $-P$ or $-N$.

It is convenient to tabulate these four possibilities as a two-by-two table.

	True State	
Test Result	+	−
P	A	B
N	C	D

The table has one cell for each of the four possibilities — each object falls into one of the cells and the table tells how many out of the set of objects falls into each cell — each cell contains a count. We have labelled these counts with letters, A, B, C, D, so we can refer to them symbolically. Completely equivalent formulations are to give the fraction or percentages of objects. (We'll work with counts in these notes.)

Watch out: a percentage can look like a count. Sometimes fractions are row or column fractions. You can easily find out by looking to see whether all 4 cells add up to 100% or whether each row (or column) on its own adds up to 100%.

The central fact that makes performance demonstrations hard to evaluate is this:

> There are two distinct ways to be wrong. Symbolically, these are $-P$ and $+N$, and called **false positives** and **false negatives**, respectively.

The word "false" refers to the fact that the test result disagrees with the true state. The words "positive" or "negative" refer to the test result itself.

In the setting of breast cancer, for example, a false positive outcome is when the test indicates the subject has the disease but this is wrong. A false negative is when the test *fails* to indicate the disease in a diseased subject.

The performance demonstration always involves four numbers, one for each cell in the two-by-two table. But these numbers aren't in a format that is directly useful. In describing a performance demonstration some other arrangement of the data — always based on the four numbers — are used. Each arrangement is useful for it's own purposes.

We will emphasize three technical terms: make sure that you know the meanings of these:

Sensitivity This is a probability called a **conditional probability**. The sensitivity answers the question: for those objects whose state is +, what is the probability of a P test result. It is calculated with $A/(A+C)$. The number is always between 0 and 1 with 1 being the best.

Specificity A different conditional probability. For those objects where the state is −, what is the probability of a N test result. Calculated with $D/(B+D)$ The number is always between 0 and 1 with 1 being the best.

Prevalence The fraction of the objects with a + state. Since the total number of objects is $n = A+B+C+D$, the prevalence is $(A+C)/n$.

Note that sensitivity and specificity refer to two disjoint sub-populations: sensitivity to the sub-population with state +, specificity to the sub-population with state −.

Two terms are quite commonly used; it is important that you know what these are:

False Positive Rate The probability (not conditional) that a randomly selected object from the demonstration set is $-P$. The number is always between 0 and 1; a zero is the best.

False Negative Rate The probability (not conditional) that a randomly selected object is $+N$. The number is always between 0 and 1; a zero is the best.

Some other terms that are sometimes used:

Accuracy $(A+D)/n$. Note that a test with high accuracy is not necessarily better than another test with low accuracy, which makes this quantity problematical. Don't be mislead by the friendly-sounding term "accuracy."

Positive Predictivity answers the question important to someone who has just had bad news from a medical test. Given that you have a positive test, what is the probability that you actually have the condition? $A/(A+B)$

Negative Predictivity Given that you have a negative test, what is the probability that you do not have the condition? $D/(C+D)$

A Canonical Description

Whatever the arrangement of the performance demonstration data, a complete description of the demonstration involves always 4 numbers. Different disciplines may use different arrangements. For us, we will set as a standard, canonical description

1 & 2 Two numbers describing the test itself: these are sensitivity and specificity.

3 A number describing the demonstration population: the prevalence.

4 A number describing the demonstration itself: the total size n of the demonstration population.

Relevance and Prevalence

Many performance demonstrations, for reasons of efficiency, are contrived to have a prevalence that is quite different from that of the population to which the test will eventually be applied. For example, in one study of a lie-detector system, the objects in the performance demonstration set were people, half of whom had been asked to carry out a mock violent act [**?**] and then to lie about it under test conditions. There is no reason to think that half of all the people who would undergo a lie detector test are lying; probably the prevalence is much less than this. In many medical performance demonstrations the prevalence in the demonstration population is much higher than in the overall population.

Given a performance demonstration of a test with one prevalence in the test population, enough information is available to compute the performance of the same test in a population with a different prevalence. The sensitivity and specificity are independent of the prevalence. In contrast, false-positive and false-negative rates are closely tied to the prevalence; these rates can be misleading if applied to a population with a different prevalence.

Another important issue is relevance; whether the demonstration objects are similar enough to the objects to whom the test will eventually be applied that the demonstration is informative. For example, prostate cancer has a very high prevalence (approaching 50%) in elderly men; tests that perform well on elderly men may not give meaningful in young or middle-aged men since the causes or preconditions of prostate cancer in the two groups may differ. Judging relevance is not a mathematical issue, but involves an understanding of the mechanism of the test. For instance, in medical tests the relevance can depend on detailed aspects of biochemistry, physiology, and genetics.

Comparing Tests

The performance of a test requires a two-number description: e.g. sensitivity and specificity. There is a trade-off between these two numbers: we can always make one of the closer to 100% by making the other one farther from 100%.

For any detection test, there is always a way to make the sensitivity 100%; simply arrange the test to always give P as an outcome. Similarly, we can always arrange to have a specificity of 100% just by always giving N as the outcome. However, we cannot always do both these things at the same time. Thus the trade-off.

To move a test's sensitivity closer to 100%, we only need a randomization device such as a spinner or the toss of a die or a coin flip. The procedure is to conduct the test in the ordinary way, then randomly construct the outcome of the test to be the indicated outcome or P. Similarly, we can move the specificity closer to 100% by randomly giving N as the outcome of the test.

Consider two possible test systems, e.g., breast self-examination vs mammography. Let's call these A and B. Test A is **unambiguously better** than test B if both the sensitivity and the specificity of A are higher than for B.

In cases where the sensitivity of A is higher than B but the specificity of A is lower than B, or vice versa, then the two tests can still be compared, but ambiguously. In some situations, it is possible — using a randomization device — to modify one of the tests to produce another test that is unambiguously better. This is described in THE DIAGRAM.

Ambiguous Comparisons and The Loss

In many situations, high sensitivity is much more valuable than specificity, or vice versa. For example, in screening donated blood for HIV, it is much more important to have a high sensitivity than a high specificity; it's very important to avoid false negatives that contaminate the blood supply and much less important to avoid the waste of donated blood that results from a false positive. Medical screening tests are generally arranged to have a high sensitivity and — due to the trade-off between sensitivity and specificity — a low specificity. This is because it is deemed more important to avoid false negatives than to avoid false positives. Much of the controversy in the medical screening literature results around the cost of a false positive — doctors usually assume this is without cost.

Whenever it is possible to assign a definite cost to a false positive and another cost to a false negative, the entire performance demonstration can be summarized by a single number called the **loss**. In such cases, it is always possible to make an unambiguous comparison of two tests: the one with the lower cost is the better one.

The Gold Standard

In many situations the true state is unknowable. For example, it's impractical to recruit actual terrorists to conduct a performance demonstration of an airport security system; mock terrorists may not provide a relevant population for testing. In medical screening, the goal is often to detect the early stages of a disease — so early that there are no definitive means of saying whether the disease is actually present.

In the absence of a definitive test, one often relies on a **gold standard**. Although the name might suggest that this is a definitive test, that is incorrect; a gold standard refers to a practice that is the best conventional one. "Best" may not be very good at all. Nonetheless, the result of the gold standard test is — for wont of a definitive test — taken to be the definition of the true state.

When comparing a new test to a gold standard, the new test will always look inferior. This is because the gold standard, by definition, has a 100% sensitivity and 100% specificity. This is a tautology: the test result (that is, the result of the gold standard test) always matches the defined true state (that is, the result of the gold standard test).

There leads unfortunately to a bias against the introduction of new tests.

Hypothesis Testing

A situation closely related to detection is **hypothesis testing** which will be the topic of a later Math 10 lecture. In hypothesis testing the analogy to the two states are called the Null Hypothesis (+) and the Alternative Hypothesis (−). The prevalence is unknowable — there is no way to measure the true state. Statisticians use the term **significance** to refer to 1 minus the sensitivity. The statistical word for specificity is **power**. I note these correspondences only to highlight the conceptual analogies between detection and hypothesis testing — sensitivity and specificity are not a conventional part of the vocabulary of hypothesis testing.

Exercise: The specificity of an ELISA test for HIV is roughly 98%. Assuming that the sensitivity is 100%, and taking the prevalence of HIV infection in low-risk groups to be 2 in 10000, what is the false positive rate in low-risk groups? Ans: 20 in 10000. If a low-risk person has a positive ELISA test, what is the probability that he has HIV? Answer: 2 in 22. Answer the same questions assuming the prevalence is 50%.

Exercise on computing false positive and false negative rates for a different prevalence.

Exercise: In criminal trials in the US, "guilty beyond a reasonable doubt" is the rule. Yet there are known cases of false positives, that is, people being wrongly convicted. State what you think would be a reasonable sensitivity and specificity of the criminal trial system. What is the difficulty of requiring 100% sensitivity? What is the difficulty of requiring 100% specificity?